

KISHOR KUMAR PAROI

AI Systems Architect | (Dev/ML/LLM/AI)Ops | Founder@KoomAI | FDE(LangGraph, Kubernetes, Kubeflow, GCP, Hybrid RAG<35ms, Agent Mesh, AI Security) | vLLM, Finetune | Data Engineer(Azure, AWS, Databricks, Snowflake) | Rust SIMD

Founder @ KoomAI (Live: koom.kishor.tech | GitHub) • B.Sc. in CSE @ RUET • Dhaka, Bangladesh / Global Remote

1703053@student.ruet.ac.bd | kishor.ruet.cse@gmail.com | +880 1724-714156 | koom.kishor.tech | Kishor.tech | macos-portfolio | 3d | hacker | linkedin | github | x.com/Mr_paroi

EDUCATION

Year	Degree / Certificate	Institute / Board	CPI / Marks	Merit Rank
2018 -- 2023	B.Sc. in Computer Science & Engineering	Rajshahi Univ. of Eng. & Tech. (RUET)	3.25 / 4.00 (D.O)	Rank 483rd in Bangladesh
2015 -- 2017	Higher Secondary Certificate (HSC)	Saint Joseph Higher Secondary School (Dhaka)	5.00 / 5.00	Rank 85th in Dhaka Board
2013 -- 2015	Secondary School Certificate (SSC)	S.J Govt Model High School (Gopalganj)	5.00 / 5.00	Rank 90th in Dhaka Board

WORK EXPERIENCE

phi0.ai — *July 2026 – Oct 2026*

Forward Deployed AI Engineer (Production RAG, GenAI & MLOps, Europe | Remote)

- Orchestrated forward deployment of **5 concurrent autonomous AI agents** executing multi-step tools via **LangGraph**; attained **99.4% task completion**, accelerating client onboarding by **4x**.
- Devised high-throughput **Hybrid RAG** with HNSW, BM25, and multi-stage re-ranking (**FlashRank + Cohere v3.5**) under **<35ms semantic cache**; curtailed hallucinations by **46%** and saved **$18K/mo**.
- Codified continuous **MLOps & GenAI flywheel** with NeMo Guardrails and Canary K8s deployments; spearheaded enterprise AI transformation, securing **$1.4M in client contracts**.

Samsung AI Core / Care Health — *March 2026 – June 2026*

Data Engineering & AI Systems Intern

- Formulated batch & streaming pipelines via **AWS S3, EMR, Athena, PySpark**; slashed operational compute overhead by **$4.2K/run (34% savings)** through asynchronous event-driven Lambda/SQS dispatch.
- Accelerated 8B/70B model inference on H100 GPUs via **FlashAttention-2** and fused kernels; delivered **3.8x throughput** on >32K context windows, mitigating p99 latency by **54%**.
- Deployed real-time health insurance claim fraud detection pipeline over **AWS Kinesis and OpenTelemetry**; eradicated false-positive claim investigations by **52%**, preserving **$380K annually**.

Independent AI Specialization & Sabbatical — *Oct 2025 – Feb 2026*

Full-Time Technical Deep Dive: Math, ML/DL, GenAI, Big Data & MLOps

- Mastered advanced **Linear Algebra** (SVD, PCA), **Calculus**, and core ML; architected Transformers, Self-Attention, and CNNs from first principles in **PyTorch**.
- Pioneered **Hybrid RAG**, **LangGraph agents**, **Big Data Lakehouses** (PySpark Delta Lake, Snowflake, Kafka streams, dbt Core), and MLOps flywheels (**MLflow, Feast, K8s**).

kepi.ai — *Jan 2022 – Sept 2025*

Senior Full Stack Engineer (3 yrs 9 mos, Remote)

- Architected zero-downtime microservices migration to **Kubernetes on AWS**; scaled platform throughput by **300%** while curtailing p95 latency by **48%** at 99.99% availability.
- Spearheaded enterprise AI initiatives: delivered autonomous multi-agent pipelines and scalable LLM microservices, driving a **22% increase in customer LTV**.
- Fortified zero-trust mesh infrastructure with **Istio mTLS and fine-grained RBAC**; precluded unencrypted service-to-service transit and diminished cross-zone egress costs by **38%**.

FOUNDER VENTURE: KOOMAI

KoomAI — Enterprise Asynchronous Video Intelligence Platform — *Jan 2025 – Dec 2025*

Founder & Principal Architect | Orchestrating Desktop, Chrome Extension, Backend & Full-Stack Next.js with AWS, CDN

Next.js | Desktop | Extension | Backend

- **WebRTC 60 FPS Engine & Desktop Client:** Devised sub-120ms real-time WebRTC media pipeline and Electron desktop client with native NVENC encoding; sustained 60 FPS 1080p across **50K+ active video minutes** (-45% CPU).
- **Autonomous 20-Agent Mesh & Video Chat Copilot:** Deployed 20 concurrent **LangGraph** agents (**<35ms p99**) and interactive video copilot generating 1-click GitHub PRs, Jira user stories, and Playwright tests.
- **NotebookLLM Studio & AI Job Recon:** Architected conversational two-host audio podcast studio, system design interview drills, and automated **ai-job-search** Trojan Horse proposals with tailored ATS CVs.
- **Pomodoro Chrome MV3 & Client Invoicing:** Engineered screen recorder with **Pomodoro focus sprints** and 432Hz meditation audio; coupled video deliverables with instant **Stripe client invoicing** ($24K MRR).

ACHIEVEMENTS & HACKATHONS

- **Competitive Programming & DSA: Ranked 232nd** in Codeforces Div-2 (**Blue Coder**), **5-Star** in CodeChef, **17th in IUPC**; Qualified **Meta Hacker Cup R3**; 2000+ DSA, participated system design hackathon.
- **Champion | Technobeast AI Agents Hackathon:** Architected Qdrant vector optimization with negative embeddings and trajectory matching.
- **7th Rank | AI Security Hackathon:** Engineered real-time prompt injection firewall (**97.2% accuracy**, **<30ms** latency).
- **National Merit Scholarships:** Conferred merit scholarships for 85th Rank in Dhaka Board (HSC) and 90th Rank in Dhaka Board (SSC).

CERTIFICATIONS & KEY COURSES

- **DeepLearning.AI / Stanford (Andrew Ng):** ML & Deep Learning Specializations (Neural Nets, CNNs, Transformers, Vector Embeddings).
- **Distributed Systems, Big Data & AI Architecture:** High-Throughput Systems & AI Architecture (Arpit Bhayani, Gaurav Sen, ByteByteGo), Big Data Engineering & Lakehouses (Mayank Agarwal, Zach Wilson), MLOps (Krish Naik), LeetCode DSA.
- **Core CS & Engineering (RUET):** Advanced Data Structures & Algorithms, Object-Oriented Programming (C++/Java), Operating Systems & System Programming, Database Systems (DBMS), Distributed Systems & Cloud Computing, Computer Networks & Internet Protocols, Theory of Computation & Automata, Compiler Design, Software Engineering & System Analysis, Computer Architecture, Artificial Intelligence, Microprocessors & Interfacing, Discrete Mathematics, and Network Security & Cryptography.

LEADERSHIP & MENTORING

- **In-Person AI Agents Masterclass:** Conducted hands-on in-person AI Agents building workshops on **LangGraph state machines**, tool calling, and production multi-agent deployment.
- **Technical Mentorship & Community Impact:** Mentored **100+ students & engineers**

TECHNICAL PROJECTS

Deploy: curl | bash

Zomato AI Lakehouse & Multimodal Platform [GitHub] [Deploy] — *Sept 2026 – Present*

Lakehouse (35M+ Rows), Feast Store & CV Engine

- Architected enterprise Lakehouse on **Snowflake & PySpark Delta Lake** processing **35M+ orders (2.3 GB)** with **dbt Core SCD-2**; curtailed warehouse compute by **42% ($18K/mo savings)**.
- Deployed real-time feature streaming via **Airflow**, **Debezium CDC WAL**, and **Feast Store** over Redis; lowered serving latency to **<5ms**, unlocking **+$340K quarterly revenue lift**.
- Operationalized multimodal CV dispute arbitration (**dHash** anti-fraud, receipt OCR) with **G-Eval CoT**; slashed fraudulent claims by **68%**, safeguarding **$1.2M in annual partner margins**.
- **Compaction & Disaster Recovery:** Tuned Delta Lake compaction, Bloom filter indexes, and automated MLflow shadow canary promotion with multi-region failover (**RPO < 60s, RTO < 4.5m**).

Hybrid Anime Intelligence & LLMOps Platform [GitHub] [Deploy] — *Aug 2026 – Present*

Polyglot Mesh (Go/Rust), RSLoRA Flywheel & Multi-LoRA vLLM

- Polyglot mesh: **Go API Gateway**, **Rust SIMD** vector scoring, and **FastAPI** AI core over **Kafka**; curtailed scoring latency from 25ms to **<0.5ms (50x speedup)**, saving **64% CPU cost**.
- Formulated hybrid retrieval combining **Dual-Tower NCF**, **ChromaDB HNSW**, and **BM25** via Reciprocal Rank Fusion (**NDCG@10 = 0.94**); drove a **28% lift in user session retention**.
- Served adapters via **vLLM Multi-LoRA dynamic multiplexing** on 1x GPU with **RSLoRA flywheel**; slashed GPU spend by **68%** while scaling to **3.4x higher token throughput**.
- **Inference Optimization:** Embedded **FlashRank ONNX local re-ranking** (+21.4% MRR), **KEDA Kafka lag autoscaling** (2 to 32 pods in 35s), and 4-bit AWQ serving 420 concurrent users/GPU.

Nexus-AI Multi-Cloud Lakehouse & Autonomous Agent Mesh [GitHub] [Deploy] — *Oct 2026 – Present*

Multi-Cloud (AWS/GCP/Azure), LangGraph, PySpark Delta Lake & Next.js 16

- Orchestrated autonomous multi-agent mesh in **LangGraph** (Data NL2SQL, BI Analyst, MLOps Health, OWASP Guardrails); automated multi-step analytical reasoning with **99.2% task reliability**.
- Engineered multi-cloud Lakehouse over **AWS S3, GCP GCS, Azure ADLS Gen2** via **PySpark Delta Lake & dbt Core**; consolidated 18M+ cross-cloud events under **<250ms** query latency.
- Built polyglot microservice gateway in **Go 1.22** (:3000), **FastAPI** core (:8081), and **Next.js 16** command cockpit with real-time SSE streaming; trimmed API routing overhead by **65%**.
- Integrated continuous ML training and model registry with **MLflow, Feast Feature Store, and Docker/Kubernetes**; reduced model deployment turnaround from days to **<12 minutes**.
- Enforced zero-trust defense via **NVIDIA NeMo Guardrails, BotShield WAF, and Terraform IaC**; halted **98%+ prompt injections** while ensuring reproducible 1-click cloud bootstrap.

Enterprise AI Security & Agentic RAG Platform [GitHub] [Deploy] — *Oct 2026 – Present*

Defense-in-Depth AI Gateway, NVIDIA NeMo & LangGraph

- Built zero-trust defense with **BotShield WAF** and **NVIDIA NeMo Guardrails**; blocked **95%+ of prompt injections & jailbreaks** (benchmarked via **Microsoft PyRIT 0.14.0**).
- Formulated stateful agentic core in **LangGraph** with MemorySaver checkpoints and **Portkey Gateway**; saved **$8.4K/month (42%)** via **<50ms semantic similarity caching**.
- Standardized tool execution via **FastMCP** with **Human-in-the-Loop (HITL)** gates and Logfire/LangSmith tracing; achieved **100% audit compliance** with zero unauthorized tool runs.
- Enforced dynamic context budgeting (-36% tokens), cryptographic HMAC-SHA256 tool invocation verification, and automated RAGAS synthetic regression evaluation harness.

Smart Manufacturing Industrial AI & MLOps [GitHub] [Deploy] — *Oct 2026 – Present*

Polyglot (Rust/Go/Next.js), PyTorch Autoencoders, GKE/K3s & ArgoCD

- Deployed IoT anomaly detection (**PyTorch Autoencoders**) and defect inspection (**CNN**, <20ms); mitigated part escape defect rate by **31%**, saving **$420K annually in scrap waste**.
- Built polyglot edge architecture: **Rust SIMD** vibration pre-filter (:8082), **Go edge gateway** (:8085), **Next.js 15** cockpit (:3000), and **PyTorch AI core** (:8000).
- Orchestrated declarative CI/CD via **Jenkins, SonarQube, Nexus 3, and ArgoCD GitOps**; curtailed deployment cycles by **96% (4h to 8m)** with zero-downtime failover.
- Integrated TensorRT INT8 edge engine (<8.2ms), non-blocking OPC-UA protocol adapter (50K tags/sec), and bidirectional gRPC edge-to-cloud streaming (-68% bandwidth).

Core Systems: Rust Git Engine & POSIX Terminal Shell [GitHub] — *August 2024*

Bare-Metal Systems in Rust, Concurrency & Pipelines

- Architected Git internals in **Rust** (SHA-1 object graph, streaming zlib SIMD decompression); matched C speed with **100% compile memory safety**.
- Engineered async POSIX terminal shell with signal multiplexing and I/O redirection; attained **sub-microsecond (<1µs) latency** with **73% lower heap allocation**.

TECHNICAL SKILLS

Languages & Frameworks: Python, C/C++, Rust (Actix), Go (Gin), SQL, Bash, TypeScript, Java (Spring Boot), FastAPI, Express, NestJS, Next.js 16, Streamlit

AI, ML & LLMOps: PyTorch, vLLM, LangGraph, LangChain, HuggingFace, TRL, PEFT (RSLoRA, QLoRA), NeFTune, MLflow, Feast, Kubeflow, AWS SageMaker, KServe, DVC

Big Data & Lakehouse: Snowflake, Databricks, Azure Data Factory (ADF), Synapse CETAS, ADLS Gen2, dbt Core, PySpark, Delta Lake, Kafka, Airflow, Debezium CDC